ILLUSTRATIVE EXAMPLES OF PRINCIPAL COMPONENT ANALYSIS USING GENSTAT/PCP

W. T. Federer, Z. D. Feng and C. E. McCulloch

BU-~~918~~-M

November 1986

ABSTRACT

In order to provide a deeper understanding of the workings of principal components, four data sets were constructed by taking linear combinations of values of two uncorrelated variables to form the X-variates for the principal components analysis. The examples highlight some of the properties and limitations of principal component analysis.

This is part of a continuing project that produces annotated computer output for principal components analysis. The complete project will involve processing four examples on SAS/PRINCOMP, BMDP/4M, SPSS-X/FACTOR, GENSTAT / PCP, and SYSTAT / FACTOR. We show here the results from GENSTAT/PCP, Version 4.04.

---

* Supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University.

1. INTRODUCTION

Principal components is a form of multivariate statistical analysis and is one method of studying the correlation or covariance structure in a set of measurements on m variables for n observations. For example, a data set may consist of n = 260 samples and m = 15 different fatty acid variables. It may be advantageous to study the structure of the 15 fatty acid variables since some or all of the variables may be measuring the same response. One simple method of studying the correlation structure is to compute the m(m-1)/2 pairwise correlations and note which correlations are close to unity. When a group of variables are all highly inter-correlated, one may be selected for use and the others discarded or the sum of all the variables may be used. When the structure is more complex, the method of principal components analysis (PCA) becomes useful.

In order to use and interpret a principal components analysis there needs to be some practical meaning associated with the various principal components. In Section 2 we describe the basic features of principal components and in Section 3 we examine some constructed examples using GENSTAT/PCP to illustrate the interpretations that are possible. In Section 4 we summarize our results.

2. BASIC FEATURES OF PRINCIPAL COMPONENT ANALYSIS

PCA can be performed on either the variances and covariances among the m variables or their correlations. One should always

check which is being used in a particular computer package program. GENSTAT can use either the X'X (where X_i is the deviation of original observation from its mean), i.e. variance-covariance multiplied by degree of freedom, or the variances and covariances or the correlations but uses the correlations by default. First we will consider analyses using the matrix of variances and covariances. A PCA generates m new variables, the principal components (PCs), by forming linear combinations of the original variables, $X = (X_1, X_2, \ldots, X_m)$, as follows:

$$PC_1 = b_{11}X_1 + b_{12}X_2 + \ldots + b_{1m}X_m = Xb_1$$

$$PC_2 = b_{21}X_1 + b_{22}X_2 + \ldots + b_{2m}X_m = Xb_2$$

$$\vdots$$

$$PC_m = b_{m1}X_1 + b_{m2}X_2 + \ldots + b_{mm}X_m = Xb_m \quad ,$$

where X_i have mean zero. In matrix notation,

$$P = (PC_1, PC_2, \ldots, PC_m) = X\,(b_1, b_2, \ldots, b_m) = XB,$$

and conversely $X = P\,B^{-1}$.

The rationale in the selection of the coefficients, b_{ij}, that define the linear combinations that are the PC_s is to try to capture as much of the variation in the original variables with as few PCs as possible. Since the variance of a linear combination of the Xs can be made arbitrarily large by selecting very large coefficients, the b_{ij} are constrained by convention so that the sum of squares of the coefficients for any PC is unity:

$$\Sigma_{j=1}^{m} b_{ij}^2 = 1 \qquad i = 1,2,\ldots,m \quad .$$

Under this constraint, the b_{1j} in PC_1 are chosen so that PC_1 has maximal variance.

If we denote the variance of X_i by s_i^2 and if we define the total variance, as $T = \Sigma_{i=1}^{m} s_i^2$, then the proportion of the variance in the original variables that is captured in PC_1 can be quantified as $var(PC_1)/T$. In selecting the coefficients for PC_2, they are further constrained by the requirement that PC_2 be uncorrelated with PC_1. Subject to this constraint and the constraint that the squared coefficients sum to one, the coefficients b_{2j} are selected so as to maximize $var(PC_2)$. Further coefficients and PCs are selected in a similar manner, by requiring that a PC be uncorrelated with all PCs previously selected and then selecting the coefficients to maximize variance. In this manner, all the PCs are constructed so that they are uncorrelated and so that the first few PCs capture as much variance as possible. The coefficients also have the following interpretation which helps to relate the PCs back to the original variables. The correlation between the i^{th} PC and the j^{th} variable is

$$b_{ij}\sqrt{var(PC_i)}/s_j \quad .$$

After all m PCs have been constructed, the following identity holds:

$$var(PC_1) + var(PC_2) + \ldots + var(PC_m) = T = \Sigma_{i=1}^{m} s_i^2 \quad .$$

This equation has the interpretation that the PCs divide up the

total variance of the Xs completely. It may happen that one or more of the last few PCs have variance zero. In such a case, all the variation in the data can be captured by fewer than m variables. Actually, a much stronger result is also true; the PCs can also be used to reproduce the actual values of the Xs, not just their variance. We will demonstrate this more explicitly later.

The above properties of PCA are related to a matrix analysis of the variance-covariance matrix of the Xs, S_x. Let D be a diagonal matrix with entries being the eigenvalues, λ_i , of S_x arranged in order from largest to smallest. Then the following properties hold:

(i) $\lambda_i = \text{var}(PC_i)$

(ii) $\text{trace}(S_x) = \Sigma_{i=1}^{m} s_i^2 = T = \Sigma_{i=1}^{m} \lambda_i = \Sigma_{i=1}^{m} \text{var}(PC_i)$

(iii) $\text{corr}(PC_i, X_j) = \dfrac{b_{ij}\sqrt{\lambda}_i}{s_j}$

(iv) $S_x = B'DB$.

The statements made above are for the case when the analysis is performed on the variance-covariance matrix of the Xs. The correlation matrix could also be used, which is equivalent to performing a PCA on the variance-covariance matrix of the standardized variables,

$$Y_i = \frac{X_i - \bar{X}_i}{s_i}$$

PCA using the correlation martrix is different in these respects:

(i) The total "variance" is just the number of variables, m.

(ii) The correlation between PC_i and X_j is given by $b_{ij}\sqrt{var(PC_i)} = b_{ij}\sqrt{\lambda_i}$. Thus PC_i is most highly correlated with the X_j having the largest coefficient in PC_i in absolute value

The experimenter must choose whether to use standardized (PCA on a correlation matrix) or unstandardized coefficients (PCA on a variance-covariance matrix). The latter is used when the variables are measured on a comparable basis. This usually means that the variables must be in the same units and have roughly comparable variances. If the variables are measured in different units then the analysis will usually be performed on the standardized scale, otherwise the analysis may only reflect the different scales of measurement. For example, if a number of fatty acid analyses are made, but the variances, s_i^2, and means, $\bar{X}_i$, are obtained on different bases and by different methods, then standardized variables could be used (PCA on the correlation matrix). The situation for using X'X is the same as the one for using variance-covariance matrix. To illustrate some of the above ideas, a number of examples have been constructed and these are described in Section 3. In each case, two variables, Z_1 and Z_2, which are uncorrelated, are used to construct X_i. Thus, all the variance can be captured with two variables and hence only two of the PCs will have nonzero variances. In matrix analysis terms, only two eigenvalues will be nonzero. An important thing to note is that in general, PCA will not recover the original

variables Z_1 and Z_2. Both standardized and nonstandardized computations will be made. In Example 3, PCA3, using X'X is also illustrated.

3. EXAMPLES

Throughout the examples we will use the variables Z_1 and Z_2 (with n = 11) from which we will construct $X_1, X_2, \ldots, X_m$. We will perform PCA on the Xs. Thus, in our constructed examples, there will only really be two underlying variables.

Values of Z_1 and Z_2

Z_1	-5	-4	-3	-2	-1	0	1	2	3	4	5
Z_2	15	6	-1	-6	-9	-10	-9	-6	-1	6	15

Notice that Z_1 exhibits a linear trend through the 11 samples and Z_2 exhibits a quadratic trend. They are also chosen to have mean zero and be uncorrelated. Z_1 and Z_2 have the following variance-covariance matrix (a variance-covariance matrix has the variance for the i^{th} variable in the i^{th} row and i^{th} column and the covariance between the i^{th} variable and the j^{th} variable in the i^{th} row and j^{th} column).

Variance-covariance matrix of Z_1 and Z_2

$$\begin{bmatrix} 11 & 0 \\ 0 & 85.8 \end{bmatrix}$$

Thus the variance of Z_1 is 11 and the covariance between Z_1 and Z_2 is zero. Also the total variance is 11 + 85.8 = 96.8.

Example 1: In this first example we analyze Z_1 and Z_2 as if they were the data. If PCA is performed on the variance-covariance matrix then the GENSTAT output is as follows (GENSTAT control language for this example and all subsequent examples is in the appendix and the boldface print was typed on computer output to explain the calculation performed):

```
PCA1: USING VARIANCE-COVARIANCE MATRIX (UNSTANDARDIZED VARIABLES)
                    PRINCIPAL COMPONENT ANALYSIS
X1        X2

-5        15
-4         6
-3        -1
-2        -6
-1        -9
 0       -10
 1        -9
 2        -6
 3        -1
 4         6
 5        15

       11 OBSERVATIONS
        2 VARIABLES

                  Covariances and Means Matrices

                        S

X1          11.0000 = S11 = S1**2                    Note:  Sij = Sji
X2          -0.0000 = S21  85.8000 = S22 = S2**2
MEAN        -0.0000 = Z1bar/n-1  -0.0000 = Z2bar/n-1  1.1000   1.1x10 = 11 = number of
                                                  (n/n-1)            observations
                        1              2              3
```

RINCIPAL COMPONENTS ANALYSIS

LATENT ROOTS

$(\lambda_1 = s_1^2)$ 1	$(\lambda_2 = s_2^2)$ 2
85.80000	11.00000

PERCENTAGE VARIANCE

1	2
88.6364	11.3636

LATENT VECTORS(LOADINGS)

	1($\underset{\sim}{b}_1$)	2($\underset{\sim}{b}_2$)
X1	b_{11} = -0.0000	b_{21} = 1.0000
X2	b_{12} = 1.0000	b_{22} = 0.0000

TRACE = 96.8000 = $\Sigma_{i=1}^{m} \lambda_i$ = T = 85.8 + 11.0 = 96.8

SIGNIFICANCE TESTS FOR EQUALITY OF REMAINING ROOTS

The following test comes from Lawley, D.N. and Maxwell, A.E. "Factor Analysis as a Statistical Method" 2nd edition (1971).

DF = $\frac{1}{2}(m-k+2)(m-k-1)$ for k=0, the χ^2 statistics is computed by the formula:

$[n-1-(\frac{1}{6})(2p+1+\frac{2}{p})][-\log_e|S| + p \log_e(\text{tr } S/p)]$

where n is the number of observations, p is the number of variables, S is variance and covariance matrix. The formula for χ^2 statistics are different when $0<k<p-1$ or when the correlation matrix is used. The interested readers may refer to p. 20-22 of Lawley and Maxwell.

NUMBERS OF UNITS AND VARIATES DIFFER BY LESS THAN 50 SO CHI-SQUARED APPROXIMATIONS ARE POOR $(n - m = 11 - 2 = 9 < 50)$

NO. OF ROOTS EXCLUDED(k)	CHI SQ	DF
0	8.1818	$2 = \frac{1}{2}(2-0+2)(2-0-1)$

$$= [11-1-(\tfrac{1}{6})(2\cdot 2+1+\tfrac{2}{2})][-\log_e(85.8\cdot 11.0)+2\log_e(\tfrac{96.8}{2})]$$

PRINCIPAL COMPONENT SCORES

$8.1818 > \chi^2_{2,\ .05} = 5.99$ for $k = 0$ reject H_o that all λ_i's are equal.

$$PC_i = b_{i1}Z_1 + b_{i2}Z_2$$

	1(PC_1)	2(PC_2)	
1	15.0000	-5.0000	$= 1(-5) + 0(15) = -5$
2	6.0000	-4.0000	
3	-1.0000	-3.0000	
4	-6.0000	-2.0000	
5	-9.0000	-1.0000	
6	-10.0000	-0.0000	
7	-9.0000	1.0000	
8	-6.0000	2.0000	
9	-1.0000	3.0000	
10	6.0000	4.0000	
11	15.0000	5.0000	

RESIDUALS

(Residuals are distances between fitted $\hat{\underline{X}}$ and $\underline{X}$. A large residual indicates an outlier, or the residuals can indicate a systematic pattern in the remaining dimensions. In our example, these residuals are all zero within rounding error.)

	1
1	8.25906E -7
2	4.76837E -7
3	0.00000E 0
4	0.00000E 0
5	0.00000E 0
6	0.00000E 0
7	3.37175E -7
8	4.12953E -7
9	0.00000E 0
10	0.00000E 0
11	0.00000E 0

We can interpret the results as follows:

1) The first principal component is

$$PC_1 = 0 \cdot X_1 + 1 \cdot X_2 = X_2$$

2) $PC_2 = 1 \cdot X_1 + 0 \cdot X_2 = X_1$

3) $Var(PC_1) = \text{eigenvalue} = 85.8 = Var(X_2)$

4) $Var(PC_2) = \text{eigenvalue} = 11.0 = Var(X_1)$

The PCs will be the same as the Xs whenever the Xs are uncorrelated. Since X_2 has the larger variance, it becomes the first principal component.

If PCA is performed on the correlation matrix we get different results.

Correlation Matrix of Z_1 and Z_2

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A correlation matrix always has unities along its diagonal and the correlation between the i^{th} variable and the j^{th} variable in the i^{th} row and j^{th} column. PCA in GENSTAT would yield the following output:

PCA1B: USING CORRELATION MATRIX (STANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

X1	X2
-5	15
-4	6
-3	-1
-2	-6
-1	-9
0	-10
1	-9
2	-6
3	-1
4	6
5	15

PRINCIPAL COMPONENTS ANALYSIS

LATENT ROOTS

$1(\lambda_1)$	$2(\lambda_2)$
1.000000	1.000000

PERCENTAGE VARIANCE

1	2
50.0000	50.0000

LATENT VECTORS(LOADINGS) = $\underset{\sim}{b}_i$

	1	2
X1	$b_{11} = 0.6925$	$b_{21} = 0.7214$
X2	$b_{12} = -0.7214$	$b_{22} = 0.6925$

TRACE = 2.0000 = $\Sigma_{i=1}^{m} \lambda_i = m = 2$

*** SIGNIFICANCE TESTS FOR EQUALITY OF REMAINING ROOTS ***

NUMBERS OF UNITS AND VARIATES DIFFER BY LESS THAN 50 SO CHI-SQUARED APPROXIMATIONS ARE POOR

NO. OF ROOTS EXCLUDED	CHI SQ	DF	This is true since $\lambda_1 = \lambda_2 = 1$.
0	0.0000	2	$\chi^2 = 0.0$

***** PRINCIPAL COMPONENT SCORES *****

$$PC_i = b_{i1}Z_{1/S_1} + b_{i2}Z_{2/S_2}$$

	1	2
1	-2.21223	0.03398
2	-1.30251	-0.42141
3	-0.54855	-0.72727
4	0.04965	-0.88360
5	0.49209	-0.89040
6	0.77878	-0.74766
7	0.90971	-0.45539
8	0.88489	-0.01359
9	0.70431	0.57774
10	0.36797	1.31861
11	-0.12412	2.20900

$$= \frac{0.7214(-5)}{3.316625} + \frac{.6925(15)}{9.262829}$$

***** RESIDUALS *****

	1
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0

$\underset{\sim}{b}_{ij}$ (correlations and scaled means)

X1	r_{11} = 1.0000E 0		
X2	r_{21} = -1.4802E-14	r_{22} = 1.0000E 0	
MEAN	$\frac{\bar{X}_1}{n-1}$ = -5.6843E-15	$\frac{\bar{X}_2}{n-1}$ = -8.5265E-15	$\frac{n}{n-1}$ = 1.1000E 0
	1	2	3

Example 2: Let $X_1 = Z_1$, $X_2 = 2Z_1$ and $X_3 = Z_2$. If the analysis is performed on the variance-covariance matrix using GENSTAT the results are:

PCA2: USING VARIANCE-COVARIANCE MATRIX (UNSTANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

X1	X2	X3
-5	-10	15
-4	-8	6
-3	-6	-1
-2	-4	-6
-1	-2	-9
0	0	-10
1	2	-9
2	4	-6
3	6	-1
4	8	6
5	10	15

COVARIANCES

S

X1	11.0000			
X2	22.0000	44.0000		
X3	-0.0000	-0.0000	85.8000	
MEAN	-0.0000	-0.0000	-0.0000	1.1000
	1	2	3	4

PRINCIPAL COMPONENTS ANALYSIS

LATENT ROOTS (λ_i)

1	2	3
85.80000	55.00000	-0.00000

PERCENTAGE VARIANCE

1	2	3
60.9375	39.0625	-0.0000

LATENT VECTORS(LOADINGS) = $\underset{\sim}{b}_i$

NOTE: Negative coefficient for PC2

	1	2	3
X1	-0.0000	-0.4472	0.8944
X2	-0.0000	-0.8944	-0.4472
X3	1.0000	-0.0000	-0.0000

TRACE = 140.8000

SIGNIFICANCE TESTS FOR EQUALITY OF REMAINING ROOTS

NUMBERS OF UNITS AND VARIATES DIFFER BY LESS THAN 50 CHI-SQUARED APPROXIMATIONS ARE POOR

NO. OF ROOTS EXCLUDED(k)	CHI SQ	DF
0	26.9240	5
1	20.9683	2

For k>1, test the hypothesis that the M-k smallest latent roots are equal.

Here, test that λ_2 and λ_3 are equal, $20.9683 > \chi^2_{2,\ .05}$. Thus λ_2 and λ_3 are not equal.

***** PRINCIPAL COMPONENT SCORES *****

NOTE: CHI SQ Test is not reliable when at least one of latent roots is zero

	1	2	3
1	15.0000	11.1803	-0.0000
2	6.0000	8.9443	-0.0000
3	-1.0000	6.7082	0.0000
4	-6.0000	4.4721	0.0000
5	-9.0000	2.2361	0.0000
6	-10.0000	0.0000	0.0000
7	-9.0000	-2.2361	0.0000
8	-6.0000	-4.4721	0.0000
9	-1.0000	-6.7082	0.0000
10	6.0000	-8.9443	-0.0000
11	15.0000	-11.1803	-0.0000

↑

← -11.1803 = -.4472(5) - .894(10) + 0(15)

***** RESIDUALS *****

	1
1	0.00173016
2	0.00138413
3	0.00103810
4	0.00069206
5	0.00034603
6	0.00000000
7	0.00034603
8	0.00069206
9	0.00103810
10	0.00138413
11	0.00173016

Analyzing the correlation matrix gives the following results:

PCA2B: USING CORRELATION MATRIX (STANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

X1	X2	X3
-5	-10	15
-4	-8	6
-3	-6	-1
-2	-4	-6
-1	-2	-9
0	0	-10
1	2	-9
2	4	-6
3	6	-1
4	8	6
5	10	15

PRINCIPAL COMPONENTS ANALYSIS

LATENT ROOTS (λ_i)

1	2	3
2.000000	1.000000	-0.000000

PERCENTAGE VARIANCE

1	2	3
66.6667	33.3333	-0.0000

LATENT VECTORS(LOADINGS) = $\underset{\sim}{b}_i$

	1	2	3
X1	-0.7071	0.0000	0.7071
X2	-0.7071	0.0000	-0.7071
X3	0.0000	1.0000	0.0000

*** TRACE = 3.0000

*** SIGNIFICANCE TESTS FOR EQUALITY OF REMAINING ROOTS ***

NUMBERS OF UNITS AND VARIATES DIFFER BY LESS THAN 50 SO CHI-SQUARED APPROXIMATIONS ARE POOR

NO. OF ROOTS EXCLUDED(k)	CHI SQ	DF
0	0.0000	5 $= \frac{1}{2}(3+2)(3-1) = 5$
1	0.0000	2

***** PRINCIPAL COMPONENT SCORES *****

	1 = PC_1	2 = PC_2	3 = PC_3
1	2.13201	1.61938	0.00000
2	1.70561	0.64775	0.00000
3	1.27920	-0.10796	-0.00000
4	0.85280	-0.64775	-0.00000
5	0.42640	-0.97163	-0.00000
6	-0.00000	-1.07958	-0.00000
7	-0.42640	-0.97163	-0.00000
8	-0.85280	-0.64775	-0.00000
9	-1.27920	-0.10796	-0.00000
10	-1.70561	0.64775	0.00000
11	-2.13201	1.61938	0.00000

$$-2.13201 = -0.707107\left[\frac{X_1 - \bar{X}_1}{S_1}\right] - .707107\left[\frac{X_2 - \bar{X}_2}{S_2}\right] + 0\left[\frac{X_3 - \bar{X}_3}{S_3}\right]$$

$$= -.707107\left[\frac{5 - 0}{3.316625}\right] - .707107\left[\frac{10 - 0}{6.63325}\right] + 0\left[\frac{15 - 0}{9.262829}\right]$$

***** RESIDUALS *****

	1
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0

S (Correlations)

X1	1.0000			
X2	1.0000	1.0000		
X3	-0.0000	-0.0000	1.0000	
MEAN	-0.0000	-0.0000	-0.0000	1.1000
	1	2	3	4

There are several items to note in these analyses:

i) There are only two nonzero eigenvalues since given X_1 and X_3, X_2 is computed from X_1.

ii) X_3 is its own principal component since it is uncorrelated with all the other variables.

iii) The sum of the eigenvalues is the sum of the variances, i.e.,

$$11 + 44 + 85.8 = 140.8$$

and

$$1 + 1 + 1 = 3 .$$

iv) For the variance-covariance analysis, the ratio of the coefficients of X_1 and X_2 in PC_2 is the same as the ratio of the variables themselves (since $X_2 = 2X_1$).

v) Since there are only two nonzero eigenvalues, only two of the PCs have nonzero variances (are nonconstant).

vi) The coefficients help to relate the variables and the PCs. In the variance-covariance analysis,

$$\text{Corr}(PC_2, X_1) = \frac{(\text{coefficient of } X_1 \text{ in } PC_2)\sqrt{\text{var}(PC_2)}}{\sqrt{\text{var}(X_1)}}$$

$$= \frac{b_{21}\sqrt{\lambda_2}}{s_1}$$

$$= \frac{-.4472\sqrt{55}}{\sqrt{11}}$$
$$= -1 \quad .$$

In the correlation analysis,

$$\text{Corr}(PC_1, X_1) = b_{11}\sqrt{\lambda_1}$$
$$= -.707107\sqrt{2}$$
$$= -1 \quad .$$

Thus, in both these cases, the variable is perfectly correlated with the PC.

vii) The Xs can be reconstructed exactly from the PCs with nonzero eigenvalues. For example, in the variance-covariance analysis, X_3 is clearly given by PC_1 . X_1 and X_2 can be recovered via the formulas

$$X_1 = PC_2/\sqrt{5}$$

$$X_2 = 2 \cdot PC_2/\sqrt{5} \quad .$$

As a numerical example,

$$-5 = -11.180/\sqrt{5} \ .$$

Or more generally,

$$\underset{\sim}{X} = \underset{\sim}{PC}\ B^{-1}$$

where B is matrix consisting of $\underset{\sim}{b}_i$'s as columns.

Example 3: For Example 3 we use $X_1 = Z_1$, $X_2 = 2(Z_1+5)$, $X_3 = 3(Z_1+5)$ and $X_4 = Z_2$. Thus X_1, X_2 and X_3 are all created from Z_1. PCA3 and PCA3B follow the procedural calls for PCA1 and PCA2 (see

pages 37 and 38). For examples PCA3C and PCA3D, use the procedural call on pages 39 and 40. Note that this gives the correct means and not the scaled means. The analyses for the variance-covariance matrix (unstandardized analysis), correlation matrix (standardized analysis), X'X matrix and correlation matrix based on X'X matrix are given below:

PCA3: USING VARIANCE-COVARIANCE MATRIX (UNSTANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

X1	X2	X3	X4
-5	0	0	15
-4	2	3	6
-3	4	6	-1
-2	6	9	-6
-1	8	12	-9
0	10	15	-10
1	12	18	-9
2	14	21	-6
3	16	24	-1
4	18	27	6
5	20	30	15

11 OBSERVATIONS
4 VARIABLES

S (Covariances and scaled mean matrix)

X1	11.0000				
X2	22.0000	44.0000			
X3	33.0000	66.0000	99.0000		
X4	-0.0000	-0.0000	-0.0000	85.8000	
MEAN	-0.0000	1.0000	1.5000	-0.0000	1.1000

NOTE: These are scaled means ($\bar{X}_i/(n-1)$).

1	2	3	4	5

PRINCIPAL COMPONENTS ANALYSIS

LATENT ROOTS

1	2	3	4
154.0000	85.8000	-0.0000	-0.0000

PERCENTAGE VARIANCE

1	2	3	4
64.2202	35.7798	-0.0000	-0.0000

LATENT VECTORS(LOADINGS) = $\underset{\sim}{b}_i$

	1 = $\underset{\sim}{b}_1$	2 = $\underset{\sim}{b}_2$	3 = $\underset{\sim}{b}_3$	4 = $\underset{\sim}{b}_4$
X1	-0.2673	0.0000	0.9514	-0.1531
X2	-0.5345	0.0000	-0.2786	-0.7979
X3	-0.8018	0.0000	-0.1314	0.5830
X4	0.0000	1.0000	0.0000	0.0000

TRACE = 239.8000

SIGNIFICANCE TESTS FOR EQUALITY OF REMAINING ROOTS

NUMBERS OF UNITS AND VARIATES DIFFER BY LESS THAN 50 SO CHI-SQUARED APPROXIMATIONS ARE POOR

NO. OF ROOTS EXCLUDED(k)	CHI SQ	DF
0	57.9493	9 = $\frac{1}{2}(4+2)(4-1)$ = 9
1	43.3078	5
2	43.3078	2

PRINCIPAL COMPONENT SCORES

	1 = PC_1	2 = PC_2	3 = PC_3	4 = PC_4
1	3.0735	15.0000	-4.2812	0.6890
2	-0.6682	6.0000	-4.2812	0.6890
3	-4.4098	-1.0000	-4.2812	0.6890
4	-8.1515	-6.0000	-4.2812	0.6890
5	-11.8931	-9.0000	-4.2812	0.6890
6	-15.6348	-10.0000	-4.2812	0.6890
7	-19.3764	-9.0000	-4.2812	0.6890
8	-23.1181	-6.0000	-4.2812	0.6890
9	-26.8598	-1.0000	-4.2812	0.6890
10	-30.6014	6.0000	-4.2812	0.6890
11	-34.3431	15.0000	-4.2812	0.6890

***** RESIDUALS *****

	1
1	0.00000000
2	0.00000000
3	0.00000000
4	0.00083756
5	0.00149866
6	0.00198681
7	0.00240951
8	0.00279660
9	0.00316119
10	0.00351030
11	0.00384814

PCA3B: USING CORRELATION MATRIX (STANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

X1	X2	X3	X4
-5	0	0	15
-4	2	3	6
-3	4	6	-1
-2	6	9	-6
-1	8	12	-9
0	10	15	-10
1	12	18	-9
2	14	21	-6
3	16	24	-1
4	18	27	6
5	20	30	15

PRINCIPAL COMPONENTS ANALYSIS

LATENT ROOTS = λ_i

1	2	3	4
3.000000	1.000000	0.000000	-0.000000

PERCENTAGE VARIANCE

1	2	3	4
75.0000	25.0000	0.0000	-0.0000

LATENT VECTORS(LOADINGS) = $\underset{\sim}{b}_i$

	1	2	3	4
X1	-0.5774	0.0000	0.4082	0.7071
X2	-0.5774	0.0000	0.4082	-0.7071
X3	-0.5774	0.0000	-0.8165	0.0000
X4	0.0000	1.0000	-0.0000	0.0000

TRACE = 4.0000

SIGNIFICANCE TESTS FOR EQUALITY OF REMAINING ROOTS

NUMBERS OF UNITS AND VARIATES DIFFER BY LESS THAN 50 SO CHI-SQUARED APPROXIMATIONS ARE POOR

NO. OF ROOTS EXCLUDED(k)	CHI SQ	DF
0	145.5609	9
1	123.5886	5
2	123.5886	2

$5 = \frac{1}{2}(4-1+2)(4-1-1) = \frac{1}{2}(5)(2) = 5$

PRINCIPAL COMPONENT SCORES

	1	2	3	4
1	1.04447	1.61938	-0.55391	-0.95940
2	0.52223	0.64775	-0.55391	-0.95940
3	0.00000	-0.10796	-0.55391	-0.95940
4	-0.52223	-0.64775	-0.55391	-0.95940
5	-1.04447	-0.97163	-0.55391	-0.95940
6	-1.56670	-1.07958	-0.55391	-0.95940
7	-2.08893	-0.97163	-0.55391	-0.95940
8	-2.61116	-0.64775	-0.55391	-0.95940
9	-3.13340	-0.10796	-0.55391	-0.95940
10	-3.65563	0.64775	-0.55391	-0.95940
11	-4.17786	1.61938	-0.55391	-0.95940

RESIDUALS

	1
1	0.00000E 0
2	0.00000E 0
3	0.00000E 0
4	0.00000E 0
5	0.00000E 0
6	0.00000E 0
7	0.00000E 0
8	1.82173E -4
9	3.61650E -4
10	4.97970E -4
11	6.20271E -4

S (Correlation matrices and scaled means)

X1	1.0000				
X2	1.0000	1.0000			
X3	1.0000	1.0000	1.0000		
X4	-0.0000	-0.0000	-0.0000	1.0000	
MEAN	-0.0000	1.0000	1.5000	-0.0000	1.1000

(NOTE: These means are scaled.)

	1	2	3	4	5

PCA3C: USING $\underset{\sim}{Y}'\underset{\sim}{Y}$ MATRIX (UNSTANDARDIZED VARIABLES)

PRINCIPAL COMPONENT ANALYSIS

(where $Y_i = X_i - \bar{X}$)

Note: This output is from default, i.e. SSP-Structure is not specified. Only variable list is given. See the control language in the Appendix. (p.39).

X1	X2	X3	X4
-5	0	0	15
-4	2	3	6
-3	4	6	-1
-2	6	9	-6
-1	8	12	-9
0	10	15	-10
1	12	18	-9
2	14	21	-6
3	16	24	-1
4	18	27	6
5	20	30	15

PRINCIPAL COMPONENTS ANALYSIS

LATENT ROOTS

Note: Latent roots are the ones from variance-covariance matrix multiplied by d.f = 11-1 = 10

1	2	3	4
1540.000	858.000	-0.000	-0.000

PERCENTAGE VARIANCE

1	2	3	4
64.2202	35.7798	-0.0000	-0.0000

LATENT VECTORS(LOADINGS)

	1	2	3	4
X1	-0.2673	0.0000	0.9545	-0.1319
X2	-0.5345	0.0000	-0.2608	-0.8039
X3	-0.8018	0.0000	-0.1443	0.5799
X4	0.0000	1.0000	0.0000	0.0000

TRACE = 2398.0000

SIGNIFICANCE TESTS FOR EQUALITY OF REMAINING ROOTS

NUMBERS OF UNITS AND VARIATES DIFFER BY LESS THAN 50 SO CHI-SQUARED APPROXIMATIONS ARE POOR

NO. OF ROOTS EXCLUDED(k)	CHI SQ	DF
0	96.7095	9
1	78.8699	5
2	78.8699	2

$$2 = \frac{1}{2}(m-k+2)(m-k-1)$$

$$= \frac{1}{2}(4-2+2)(4-2-1) = 2$$

***** PRINCIPAL COMPONENT SCORES *****

	1	2	3	4
1	18.7083	15.0000	-0.0000	0.0000
2	14.9666	6.0000	-0.0000	0.0000
3	11.2250	-1.0000	-0.0000	0.0000
4	7.4833	-6.0000	-0.0000	0.0000
5	3.7417	-9.0000	-0.0000	0.0000
6	-0.0000	-10.0000	0.0000	-0.0000
7	-3.7417	-9.0000	0.0000	-0.0000
8	-7.4833	-6.0000	0.0000	-0.0000
9	-11.2250	-1.0000	0.0000	-0.0000
10	-14.9666	6.0000	0.0000	-0.0000
11	-18.7083	15.0000	0.0000	-0.0000

$$-18.7083 = -.2673(X_1 - \bar{X}_1) - .5345(X_2 - \bar{X}_2) - .8018(X_3 - \bar{X}_3) + 0(X_4 - \bar{X}_4)$$

$$= -.2673(5) - .5345(20 - 10) - .8018(30 - 15)$$

OR, more generally:

$$PC_i = \underset{\sim}{Y}\, \underset{\sim}{b}_i$$

$\underset{\sim}{Y}$ is a row vector of observations

***** RESIDUALS *****

	1
1	0.00140053
2	0.00112042
3	0.00084032
4	0.00056021
5	0.00028010
6	0.00000000
7	0.00028010
8	0.00056021
9	0.00084032
10	0.00112042
11	0.00140053

PCA3D: USING CORRELATION MATRIX BASED ON Y'Y

X1	X2	X3	X4
-5	0	0	15
-4	2	3	6
-3	4	6	-1
-2	6	9	-6
-1	8	12	-9
0	10	15	-10
1	12	18	-9
2	14	21	-6
3	16	24	-1
4	18	27	6
5	20	30	15

PRINCIPAL COMPONENTS ANALYSIS

LATENT ROOTS = λ_i

1	2	3	4
3.000000	1.000000	-0.000000	-0.000000

PERCENTAGE VARIANCE

1	2	3	4
75.0000	25.0000	-0.0000	-0.0000

LATENT VECTORS(LOADINGS) = $\underset{\sim}{b}_i$

	1	2	3	4
X1	-0.5774	0.0000	0.6110	0.5417
X2	-0.5774	0.0000	-0.7746	0.2583
X3	-0.5774	0.0000	0.1636	-0.7999
X4	0.0000	1.0000	0.0000	-0.0000

TRACE = 4.0000

SIGNIFICANCE TESTS FOR EQUALITY OF REMAINING ROOTS

NUMBERS OF UNITS AND VARIATES DIFFER BY LESS THAN 50 SO CHI-SQUARED APPROXIMATIONS ARE POOR

NO. OF ROOTS EXCLUDED	CHI SQ	DF
0	0.0000	9
1	0.0000	5
2	0.0000	2

PRINCIPAL COMPONENT SCORES = PC_i

	1 = PC_1	2 = PC_2	3 = PC_3	4 = PC_4
1	0.825723	0.512092	0.000000	0.000000
2	0.660578	0.204837	0.000000	0.000000
3	0.495434	-0.034139	0.000000	0.000000
4	0.330289	-0.204837	0.000000	0.000000
5	0.165145	-0.307255	0.000000	0.000000
6	-0.000000	-0.341394	0.000000	0.000000
7	-0.165145	-0.307255	-0.000000	-0.000000
8	-0.330289	-0.204837	-0.000000	-0.000000
9	-0.495434	-0.034139	-0.000000	-0.000000
10	-0.660578	0.204837	-0.000000	-0.000000
11	-0.825723	0.512092	-0.000000	-0.000000

$$-0.825723 = \frac{-.5774}{\sqrt{n-1}}\left[\frac{X_1-\bar{X}_1}{S_1}\right] - \frac{.5774}{\sqrt{n-1}}\left[\frac{X_2-\bar{X}_2}{S_2}\right] - \frac{.5774}{\sqrt{n-1}}\left[\frac{X_3-\bar{X}_2}{S_3}\right] + \frac{0}{\sqrt{n-1}}\left[\frac{X_4-\bar{X}_4}{S_4}\right]$$

$$= \frac{-.5774}{\sqrt{10}}\left[\frac{5}{3.317}\right] - \frac{.5774}{\sqrt{10}}\left[\frac{20-10}{6.633}\right] - \frac{.5774}{\sqrt{10}}\left[\frac{30-15}{9.950}\right]$$

***** RESIDUALS *****

	1
1	1.56445E -4
2	1.25156E -4
3	9.38666E -5
4	6.25778E -5
5	3.12889E -5
6	0.00000E 0
7	3.12889E -5
8	6.25778E -5
9	9.38666E -5
10	1.25155E -4
11	1.56444E -4

S (Correlation matrices and means)

X1	1.0000					
X2	1.0000	1.0000				
X3	1.0000	1.0000	1.0000			
X4	-0.0000	-0.0000	-0.0000	1.0000		
MEAN	-0.0000	10.0000	15.0000	-0.0000	11.0000	→ last entry is the number of observations

(NOTE: These are correct means.)

1	2	3	4	5

For the variance-covariance analysis, the coefficients in PC_1 are in the same ratio as their relationship to Z_1. In the correlation analysis X_1, X_2 and X_3 have equal coefficients. In both analyses, as expected, the total variance is equal to the sum of the variances for the PCs. In both cases two PCs, PC_3 and PC_4, have zero variance; in the correlation analysis the PCs are identically zero but in the variance-covariance analysis they are constant, but not zero.

Example 4. In this example we take more complicated combinations of Z_1 and Z_2.

$$X_1 = Z_1$$

$$X_2 = 2Z_1$$

$$X_3 = 3Z_1$$

$$X_4 = Z_1/2 + Z_2$$

$$X_5 = Z_1/4 + Z_2$$

$$X_6 = Z_1/8 + Z_2$$

$$X_7 = Z_2$$

Note that X_1, X_2 and X_3 are colinear (they all have correlation unity) and X_4, X_5, X_6 and X_7 have steadily decreasing correlations with X_1. The PCAs for the variance-covariance and correlation matrices are given below:

PCA4: USING VARIANCE-COVARIANCE MATRIX (UNSTANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

11 OBSERVATIONS
7 VARIABLES

S (Covariance and mean matrix)

X1	11.0000							
X2	22.0000	44.0000						
X3	33.0000	66.0000	99.0000					
X4	5.5000	11.0000	16.5000	88.5500				
X5	2.7500	5.5000	8.2500	87.1750	86.4875			
X6	1.3750	2.7500	4.1250	86.4875	86.1438	85.9719		
X7	-0.0000	-0.0000	-0.0000	85.8000	85.8000	85.8000	85.8000	
MEAN	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	1.100

1	2	3	4	5	6	7
X1	X2	X3	X4	X5	X6	X7
-5	-10	-15	12.5	13.75	14.375	15
-4	-8	-12	4.0	5.00	5.500	6
-3	-6	-9	-2.5	-1.75	-1.375	-1
-2	-4	-6	-7.0	-6.50	-6.250	-6
-1	-2	-3	-9.5	-9.25	-9.125	-9
0	0	0	-10.0	-10.00	-10.000	-10
1	2	3	-8.5	-8.75	-8.875	-9
2	4	6	-5.0	-5.50	-5.750	-6
3	6	9	0.5	-0.25	-0.625	-1
4	8	12	8.0	7.00	6.500	6
5	10	15	17.5	16.25	15.625	15

PRINCIPAL COMPONENTS ANALYSIS

LATENT ROOTS = λ_i

1	2	3	4	5	6	7
347.0151	153.7943	0.0000	0.0000	-0.0000	-0.0000	-0.0000

PERCENTAGE VARIANCE

1	2	3	4	5	6	7
69.2908	30.7092	0.0000	0.0000	-0.0000	-0.0000	-0.0000

LATENT VECTORS(LOADINGS) = $\underset{\sim}{b}_i$

	1	2	3	4	5	6	7
X1	-0.0250	-0.2648	-0.0002	-0.3530	0.8944	-0.0550	0.0406
X2	-0.0500	-0.5296	-0.0005	-0.7059	-0.4472	-0.1101	0.0811
X3	-0.0751	-0.7944	0.0915	0.5721	-0.0000	0.1636	-0.0312
X4	-0.5048	-0.0274	-0.8018	0.0416	-0.0000	-0.1178	-0.2930
X5	-0.4986	0.0388	0.2413	0.1640	0.0000	-0.6573	0.4825
X6	-0.4954	0.0719	0.5385	-0.1054	0.0000	0.0620	-0.6666
X7	-0.4923	0.1050	0.0220	-0.1002	-0.0000	0.7130	0.4772

TRACE = 500.8094

SIGNIFICANCE TESTS FOR EQUALITY OF REMAINING ROOTS

NUMBERS OF UNITS AND VARIATES DIFFER BY LESS THAN 50 SO CHI-SQUARED APPROXIMATIONS ARE POOR

NO. OF ROOTS EXCLUDED(k)	CHI SQ	DF
0	563.1576	27
1	481.1394	20
2	481.1394	14
3	481.1394	9
4	481.1394	5
5	481.1394	2

$27 = \frac{1}{2}(m-k+2)(m-k-1) = \frac{1}{2}(7-0+2)(7-0-1) = 27$

PRINCIPAL COMPONENT SCORES

	1	2	3	4	5	6	7
1	-25.9208	21.3322	-0.0000	-0.0000	0.0000	0.0000	-0.0000
2	-8.7899	15.9370	-0.0000	-0.0000	0.0000	0.0000	-0.0000
3	4.3588	10.9181	-0.0000	-0.0000	0.0000	-0.0000	0.0000
4	13.5252	6.2754	0.0000	-0.0000	0.0000	-0.0000	0.0000
5	18.7094	2.0089	0.0000	-0.0000	0.0000	-0.0000	0.0000
6	19.9113	-1.8813	0.0000	-0.0000	-0.0000	-0.0000	0.0000
7	17.1310	-5.3952	0.0000	0.0000	-0.0000	-0.0000	0.0000
8	10.3683	-8.5329	0.0000	0.0000	-0.0000	-0.0000	0.0000
9	-0.3766	-11.2943	0.0000	0.0000	-0.0000	-0.0000	0.0000
10	-15.1037	-13.6795	-0.0000	0.0000	-0.0000	0.0000	-0.0000
11	-33.8131	-15.6884	-0.0000	0.0000	-0.0000	0.0000	-0.0000

$$-0.0 = 0.8944(5) - 04472(10) - 0(15) - 0(17.5) + 0(16.25) + 0(15.625) - 0(15) = 0$$

***** RESIDUALS *****

	1
1	0.0101166
2	0.0058595
3	0.0040048
4	0.0045667
5	0.0054862
6	0.0057451
7	0.0051710
8	0.0040467
9	0.0038625
10	0.0065797
11	0.0114167

PCA4B: USING CORRELATION MATRIX (STANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

-5	-10	-15	12.5	13.75	14.375	15
-4	-8	-12	4.0	5.00	5.500	6
-3	-6	-9	-2.5	-1.75	-1.375	-1
-2	-4	-6	-7.0	-6.50	-6.250	-6
-1	-2	-3	-9.5	-9.25	-9.125	-9
0	0	0	-10.0	-10.00	-10.000	-10
1	2	3	-8.5	-8.75	-8.875	-9
2	4	6	-5.0	-5.50	-5.750	-6
3	6	9	0.5	-0.25	-0.625	-1
4	8	12	8.0	7.00	6.500	6
5	10	15	17.5	16.25	15.625	15

PRINCIPAL COMPONENTS ANALYSIS

LATENT ROOTS = λ_i

1	2	3	4	5	6	7
4.052167	2.947833	0.000000	0.000000	0.000000	0.000000	0.000000

PERCENTAGE VARIANCE

1	2	3	4	5	6	7
57.8881	42.1119	0.0000	0.0000	0.0000	0.0000	0.0000

LATENT VECTORS(LOADINGS)

	1	2	3	4	5	6	7
X1	-0.1443	-0.5573	-0.4054	0.0593	0.0266	0.0013	0.7071
X2	-0.1443	-0.5573	-0.4054	0.0593	0.0266	0.0013	-0.7071
X3	-0.1443	-0.5573	0.7876	-0.2128	0.0144	0.0529	-0.0000
X4	-0.4933	0.0683	0.1888	0.8204	-0.1900	-0.0846	0.0000
X5	-0.4863	0.1188	-0.0589	-0.3960	-0.1258	-0.7571	-0.0000
X6	-0.4813	0.1441	-0.1074	-0.3339	-0.5141	0.6002	-0.0000
X7	-0.4754	0.1692	-0.0200	-0.0796	0.8259	0.2378	-0.0000

TRACE = 7.0000

SIGNIFICANCE TESTS FOR EQUALITY OF REMAINING ROOTS ***

NUMBERS OF UNITS AND VARIATES DIFFER BY LESS THAN 50 SO CHI-SQUARED APPROXIMATIONS ARE POOR

NO. OF ROOTS EXCLUDED	CHI SQ	DF
0	1113.7754	27
1	1085.1265	20
2	330.3020	14
3	328.0259	9
4	325.2202	5
5	325.1621	2

PRINCIPAL COMPONENT SCORES = PC_i

	1	2	3	4	5	6	7
1	-2.23779	3.28436	0.00000	-0.00000	0.00000	0.00000	0.00000
2	-0.54253	2.30445	0.00000	-0.00000	0.00000	0.00000	0.00000
3	0.73683	1.43222	0.00000	-0.00000	0.00000	-0.00000	0.00000
4	1.60029	0.66767	0.00000	-0.00000	-0.00000	-0.00000	0.00000
5	2.04785	0.01080	0.00000	-0.00000	-0.00000	-0.00000	0.00000
6	2.07951	-0.53839	0.00000	-0.00000	-0.00000	-0.00000	0.00000
7	1.69526	-0.97991	-0.00000	0.00000	-0.00000	-0.00000	-0.00000
8	0.89511	-1.31374	-0.00000	0.00000	-0.00000	-0.00000	-0.00000
9	-0.32093	-1.53990	-0.00000	0.00000	-0.00000	0.00000	-0.00000
10	-1.95288	-1.65838	-0.00000	0.00000	-0.00000	0.00000	-0.00000
11	-4.00073	-1.66918	-0.00000	0.00000	0.00000	0.00000	-0.00000

$$0 = .7071\left[\frac{5-0}{3.317}\right] - 7.071\left[\frac{10-0}{6.633}\right] + 0\left[\frac{15-0}{9.950}\right] + 0\left[\frac{17.5-0}{9.410}\right] + 0\left[\frac{16.25-0}{9.300}\right] + 0\left[\frac{15.625-0}{9.272}\right] + 0\left[\frac{15-0}{9.263}\right]$$

***** RESIDUALS *****

	1
1	8.77804E -4
2	7.22562E -4
3	4.92339E -4
4	1.11615E -4
5	0.00000E 0
6	0.00000E 0
7	0.00000E 0
8	3.51122E -4
9	5.19716E -4
10	5.48122E -4
11	2.79035E -4

S (Correlation matrix)

X1	1.0000							
X2	1.0000	1.0000						
X3	1.0000	1.0000	1.0000					
X4	0.1762	0.1762	0.1762	1.0000				
X5	0.0892	0.0892	0.0892	0.9961	1.0000			
X6	0.0447	0.0447	0.0447	0.9912	0.9990	1.0000		
X7	-0.0000	-0.0000	-0.0000	0.9843	0.9960	0.9990	1.0000	
MEAN	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	1.1000

	1	2	3	4	5	6	7	8

We note several things:

i) In both analyses there are only two eigenvalues that are nonzero indicating that only two variables are needed. This is not readily apparent from the correlation or variance-covariance matrix.

ii) In PC_1, PC_2 and PC_3 where the standardizrd X_1, X_2 and X_3 are the same, they have the same coefficients.

iii) Neither PCA recovers Z_1 and Z_2. The PCAs with nonzero variances have elements of both Z_1 and Z_2 in them, i.e., neither PC_1 or PC_2 is perfectly correlated with one of the Zs.

4. SUMMARY

PCA provides a method of extracting structure from the variance-covariance or correlation matrix. If a multivariate data set is actually constructed in a linear fashion from fewer variables, then PCA will discover that structure. PCA constructs linear combinations of the original data, $\underset{\sim}{X}$, with maximal variance:

$$P = XB .$$

This relationship can be inverted to recover the Xs from the PCs (actually only those PCs with nonzero eigenvalues are needed - see example 2). Though PCA will often help discover structure in a data set, it does have limitations. It will not necessarily recover the exact underlying variables, even if they were uncorrelated (Example 4). Also, by its construction, PCA is limited to searching for linear structures in the Xs.

APPENDIX

Example 1: Control Language for PCA1

Control language is typed in upper case and comments are bolded. Refer to GENSTAT RELEASE 4.04 PART II, 1983 for program documentation.

```
'REFE' PRINONE
'UNITS' $ 11
'SET' VARIATES=X1, X2
'READ/P' X1, X2      ⇒ input variables
'RUN'
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
 -9
  -6
3 -1
4 6
5 15
'EOD'                ⇒ signals GENSTAT that it is the end of data
'PRIN/P' X1, X2, $ 9.4  ⇒ print out data
'DSSP' S $ VARIATES ⇒ specify X1 and X2 as columns of X
'SSP' S              ⇒ calculate Y'Y
'CALC' S=S/10    ⇒ divided by n-1 = (11-1) to get variance-covariance matrix
'PRINT' S $ 9.4      ⇒ print out variance-covariance matrix
'PCP / PRIN=LTRCS' VARIANCES=S ⇒ print out   L - Latent roots and vectors
'RUN'                                       T - Trace of Y'Y
'STOP'                                      R - The Residuals
                                            C - The fitted values
                                            S - Asymptotic Chi-square test
                              "variance = s" ⇒ Define variance-covariance
                                            matrix to work with
```

Example 2: Control Language for PCA2B

```
'REFE' PRINONE
'UNITS' $ 11
'SET' VARIATES=X1, X2, X3
'READ/P' X1, X3
'RUN'
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
'EOD'
'CALC' X2 = X1 * 2 ⇒ creates X1,X2,X3
'PRIN/P' X1, X2, X3 $ 9.4
'DSSP' S $ VARIATES
'SSP' S
 CALC' S=S/10
'PCP / PRIN=LTRCS, CORR=Y' VARIANCES=S; SSPCALC = S ⇒ define GENSTAT to use
'PRINT' S $ 9.4                                      correlation matrix to compute
'RUN'                                                PCA by "CORR = Y" command.
'STOP'                                               Save correlation matrix in S
                                                     for printout or other use.
```

Example 3: Control Language for PCA3D

```
'REFE' PRINONE
'UNITS' $ 11
'SET' VARIATES=X1, X2, X3, X4
'READ/P' X1, X4
'RUN'
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
'EOD'
'CALC' X2 = 2*(X1+5)
'CALC' X3 = 3*(X1+5)
'PRIN/P' X1, X2, X3, X4 $ 9.4
'DSSP' S $ VARIATES
 SSP' S
 PCP / PRIN=LTRCS, CORR=Y' VARIANCES=S; SSPCALC = S ⇒ Note that before 'PCP
'PRINT' S $ 9.4                                      command, "calc" S=S/10" is
'RUN'                                                omitted, GENSTAT works on Y'Y
'STOP'          If we omit this command              instead of variance-covariance
                we will get output PCA3C             matrix. GENSTAT then compute
                                                       correlation matrix from Y'Y.
```

Example 4: Control Language for PCA4

```
'REFE' PRINONE
'UNITS' $ 11
'SET' VARIATES=X1, X2, X3, X4, X5, X6, X7
'READ/P' X1, X7
'RUN'
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
'EOD'
'CALC' X2=2*X1
'CALC' X3=3*X1
'CALC' X4=X7+(X1/2)
'CALC' X5=X7+(X1/4)
'CALC' X6=X7+(X1/8)
'PRIN/P' X1, X2, X3, X4, X5, X6, X7 $ 9.4
'DSSP' S $ VARIATES
'SSP' S
'CALC' S=S/10
'PRINT' S $ 9.4
'PCP / PRIN=LTRCS' VARIANCES=S;
'RUN'
'STOP'
```